

Margin-Based Generalization for Quantum Machine Learning with Classical Data Encoding

Ju-Young Ryu^{1,*} Tak Hur^{2,3,*} Daniel K. Park^{2,4,5}

¹Korea Institute of Science and Technology Information, Republic of Korea

²Department of Statistics and Data Science, Yonsei University, Republic of Korea

³ Department of Applied Mathematics, University of Waterloo, Canada

⁴Department of Applied Statistics, Yonsei University, Republic of Korea

⁵Department of Quantum Information, Yonsei University, Republic of Korea

jyryu@kisti.re.kr tak.hur@uwaterloo.ca dkd.park@yonsei.ac.kr

Abstract

Understanding generalization is essential for assessing the practical viability of quantum machine learning. In many applications, quantum models operate on classical data that must first be encoded into quantum states, making the encoding part of the effective hypothesis class. We develop a margin-based generalization framework for this setting, covering both fixed and trainable quantum data encodings. For fixed encodings, the existing margin bound transfers unchanged. For trainable encodings, we derive a bound using a two-scale covering argument that combines the original margin-dependent term with an additive encoding-capacity term of order $\tilde{O}(\sqrt{L/m})$, where L is the number of trainable encoding parameters and m is the training-sample size. Across 19,440 simulated training runs spanning various image datasets, circuit architectures, and optimization settings, the inverse margin performs best among five baseline and two bound-derived complexity measures under conditional-independence and 90th-percentile sign-error criteria, within bootstrap uncertainty. These results provide a unified margin-based view of generalization in quantum machine learning with classical data.

1 Introduction

Quantum neural networks (QNNs) use parameterized quantum circuits as trainable models for learning tasks. As in classical machine learning, their practical utility depends on *generalization*: the ability to make accurate predictions on unseen data from a finite training sample. Generalization in QML has been studied through parameter counts (Caro et al., 2022), effective dimension (Abbas et al., 2021), and quantum information-theoretic quantities (Banchi et al., 2021).

Most of these analyses produce *uniform* bounds that hold simultaneously for every hypothesis in the model class. Uniform bounds have faced sustained criticism in classical deep learning. Zhang et al. (2021) showed that deep networks easily fit random labels, so any bound that applies equally to all hypotheses in the class cannot distinguish models that generalize from models that memorize. The same failure occurs in QML. Gil-Fuster et al. (2024) demonstrated that even few-qubit QNNs can perfectly fit randomized labels on a benchmark quantum dataset, rendering uniform bounds equally uninformative in the quantum setting.

Classical learning theory responded to this critique with *margin-based* analyses (Cortes and Vapnik, 1995; Bartlett et al., 2017; Neyshabur et al., 2018). These analyses replace the worst case over the hypothesis class with a data-dependent quantity that measures how confidently the trained model separates the training sample. Large-scale empirical studies subsequently confirmed that margin-based complexity measures are among the most reliable predictors of generalization in deep networks (Jiang et al., 2019; Dziugaite et al., 2020). Hur and Park (2025) recently extended this line of work to QML. They established a margin-based generalization bound for QNNs classifying *quantum* data and showed empirically, on quantum phase recognition tasks, that margin-based metrics predict generalization better than parameter-based metrics.

Near-term applications of QML, however, predominantly involve *classical* data, which must be loaded into a quantum computer by a data-encoding circuit. The encoding is known to shape the expressivity of

*These authors contributed equally to this work.

the resulting model (Schuld et al., 2021; Pérez-Salinas et al., 2020) and its generalization behavior (Caro et al., 2021). When the encoding is *fixed*, each classical input is mapped to a quantum state once and for all. The classical-data problem then reduces exactly to the quantum-data one, and the margin theory of Hur and Park (2025) transfers unchanged (Proposition 1). Modern QNN architectures, however, often make the encoding itself *trainable*, either by interleaving parameterized layers with data re-uploading blocks (Pérez-Salinas et al., 2020; Lloyd et al., 2020) or by learning the embedding with a classical neural network (Hur et al., 2024). In this case, the encoded sample matrix depends on learned parameters, so the fixed-matrix covering argument of Hur and Park (2025) no longer applies directly.

This paper makes the following theoretical and empirical contributions.

- We show that for any fixed data encoding the quantum-data margin bound holds without modification (Proposition 1). We then prove a margin-based generalization bound for QNNs whose inputs are produced by a trainable encoding with L parameters (Theorem 2). The bound separates into the same margin-normalized term $(bE/\gamma)\sqrt{n/m}$ as in the quantum-data case and an additive encoding-capacity term of order $\sqrt{L/m}$, up to logarithmic factors. The proof introduces a two-scale covering argument that covers the encoding parameter space and the QNN unitary class at separate scales. A corollary for QNNs whose trainable parameters all lie in the encoding circuit (Corollary 1) shows that the margin then enters the bound only logarithmically.
- Following the predictor-evaluation methodology of Jiang et al. (2019) and Dziugaite et al. (2020), we conduct a systematic empirical study of 19,440 QNN models trained on binary classification tasks derived from MNIST, KMNIST, and Fashion-MNIST. The study sweeps four embedding circuits, four variational ansätze, three re-uploading depths, three qubit counts, and a grid of optimization hyperparameters. Five baseline complexity measures are evaluated with a conditional independence test and a sign-error analysis. Margin-based measures consistently outperform parameter-count measures on both criteria.
- Guided by the derived bounds, we test two theorem-derived complexity measures, the composite $\sqrt{n_q/m_m} + c\sqrt{n_p}$ of Theorem 2 and $n_p \ln(n_p/m_m)$ of Corollary 1. Neither improves on the inverse margin. With c tuned in-sample, the composite lowers the sign-error only slightly, within the bootstrap uncertainty. Both measures score below the inverse margin in the conditional independence test. We attribute this to the parameter count being fixed by the architecture hyperparameters, on which the test conditions.

Together, these results extend the margin-based analysis of Hur and Park (2025) to QNNs trained on classical data and quantify the additional capacity introduced by learning the encoding.

2 Related Work

Margins and generalization in classical learning. Margin analysis dates back to support vector machines (Cortes and Vapnik, 1995) and underlies modern norm-based generalization theory for deep networks, notably the spectrally normalized margin bounds of Bartlett et al. (2017) and their PAC-Bayesian counterpart (Neyshabur et al., 2018). Beyond bounds, large-scale empirical evaluations established protocols for judging whether a complexity measure genuinely *predicts* generalization rather than merely correlating with it. Jiang et al. (2019) introduced a conditional independence criterion inspired by probabilistic causality, and Dziugaite et al. (2020) advocated distributional robustness of sign-errors across experimental environments. We adopt both protocols.

Generalization in QML. Uniform generalization bounds for QML have been derived from parameter counts (Caro et al., 2022), effective dimension (Abbas et al., 2021), and quantum information-theoretic arguments (Banchi et al., 2021). Gil-Fuster et al. (2024) showed that QNNs overfit randomized labels, questioning the practical value of uniform bounds, in analogy with Zhang et al. (2021). Hur and Park (2025) established a margin-based bound for QNNs with quantum input data by covering the image of the data matrix under

the unitary class, and demonstrated empirically that margin-based metrics dominate parameter-based ones. More recently, Rodriguez-Grasa et al. (2026) derived solution-dependent PAC-Bayesian bounds for layered quantum models. Our theory instead extends the quantum-data margin bound to classical inputs with trainable encodings, and our experiments evaluate margin-based predictors in that setting.

Data encoding. For classical data, the choice of quantum embedding determines the model’s frequency content (Schuld et al., 2021) and its generalization behavior. Caro et al. (2021) derived generalization bounds for *fixed* encodings via fat-shattering dimensions of the induced function class. Data re-uploading (Pérez-Salinas et al., 2020) interleaves encoding and trainable layers. Trainable embeddings (Lloyd et al., 2020) and neural quantum embeddings (Hur et al., 2024; Liu et al., 2025; Kim et al., 2026; Choi et al., 2025) optimize the encoding to separate the classes in Hilbert space, which enlarges margins (Hur and Park, 2025). Our results cover both regimes: a margin-based bound that carries over to fixed encodings (Proposition 1), and a bound that quantifies the capacity cost of training the encoding (Theorem 2).

3 Preliminaries

Multiclass classification with QNNs. We consider k -class classification of n -qubit quantum states, writing $N = 2^n$. A QNN applies a unitary U from an accessible class $\mathbb{U}_Q \subseteq \mathbb{U}(N)$, followed by a positive operator-valued measurement (POVM) $\{E_i\}_{i=1}^k$. For a pure input state $x \in \mathbb{C}^N$, the model outputs the vector of class probabilities

$$h(x) = (x^\dagger U^\dagger E_i U x)_{i=1}^k \in [0, 1]^k,$$

and predicts $\arg \max_i h(x)_i$. Given m i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^m$ from an unknown distribution $\mathcal{D}$, the true risk is $R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{1}(\arg \max_i h(x)_i \neq y)$.

Margins. The *margin operator* $\mathcal{M}(v, y) = v_y - \max_{i \neq y} v_i$ measures the gap between the correct class probability and the strongest competitor. The sample margin of (x, y) is $\mathcal{M}(h(x), y)$. For a margin level $\gamma > 0$, the ramp loss $\ell_\gamma : \mathbb{R} \rightarrow [0, 1]$ is defined by $\ell_\gamma(t) = \min\{1, \max\{0, 1 - t/\gamma\}\}$. The empirical margin loss is $\hat{R}_\gamma(h) = m^{-1} \sum_{i=1}^m \mathbf{1}(h(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} h(x_i)_j)$. Writing $\mathcal{F}_\gamma$ for the class of functions $(x, y) \mapsto \ell_\gamma(\mathcal{M}(h(x), y))$ induced by the hypothesis class, the standard margin bound (Mohri et al., 2018) states that, for any $0 < \delta < 1$, with probability at least $1 - \delta$, for all h ,

$$R(h) \leq \hat{R}_\gamma(h) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}, \quad (1)$$

where $\mathfrak{R}$ denotes the empirical Rademacher complexity of the loss class restricted to the sample.

Margin bound for quantum data. Two constants of the QNN control the Rademacher complexity in (1). The *measurement constant* is $E = (\sum_{i=1}^k \|E_i\|_\sigma^2)^{1/2}$, where $\|\cdot\|_\sigma$ is the spectral norm. The *unitary distance bound* b satisfies $\|U - U_{\text{ref}}\|_{2,1} \leq b$ for all $U \in \mathbb{U}_Q$ and some reference unitary U_{ref} , where $\|\cdot\|_{2,1}$ is the sum of the Euclidean norms of the columns. Covering the image $\mathbb{U}_Q X = \{UX : U \in \mathbb{U}_Q\}$ of the sample data matrix $X \in \mathbb{C}^{N \times m}$ with matrix sparsification techniques (Zhang, 2002), Hur and Park (2025) obtained the following.

Theorem 1 (Hur and Park, 2025). *Consider an n -qubit QNN with unitary class $\mathbb{U}_Q$ (distance bound b) and POVMs $\{E_i\}_{i=1}^k$ for k -class classification of quantum states. For any $\gamma > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$ over an i.i.d. sample of size m , every hypothesis h satisfies*

$$R(h) \leq \hat{R}_\gamma(h) + \tilde{O} \left(\frac{bE}{\gamma} \sqrt{\frac{n}{m}} + \sqrt{\frac{\ln(2/\delta)}{m}} \right).$$

When the k POVM elements are nonzero projectors, $E = \sqrt{k}$, so the measurement factor is common across models with the same number of classes. The remaining factors b and n must still be accounted for when comparing different unitary classes or circuit widths.

4 Margin Bounds with Classical Data Encoding

Fixed encodings. A fixed encoding transports the data distribution onto the space of quantum states, so the classical-data problem is an instance of the quantum-data one and the quantum-data bound transfers directly.

Proposition 1 (Fixed encodings). *For any fixed encoding $x \mapsto |\psi(x)\rangle$ of classical inputs into n -qubit states, the bound of Theorem 1 holds for the hypothesis class of all $h(x) = (\psi(x)^\dagger U^\dagger E_i U \psi(x))_{i=1}^k$ with $U \in \mathbb{U}_Q$.*

Proof. The encoded pairs $(\psi(x_i), y_i)$ are i.i.d. samples from the pushforward of $\mathcal{D}$ under $(x, y) \mapsto (\psi(x), y)$, and under this identification the risks, the margins, and the hypothesis class are exactly those of the quantum-data setting. $\square$

Trainable encodings. Modern encodings, however, are trained together with the classifier, and the reduction above then breaks down. The encoded data matrix is no longer fixed but moves with the encoding parameters. We model a trainable encoding as a parameterized circuit that prepares $|\psi(x; \phi)\rangle = V(x; \phi) |0\rangle^{\otimes n}$ with

$$V(x; \phi) = \left(\prod_{j=1}^L e^{iH_j \phi_j} W_j(x) \right) W_0(x), \quad (2)$$

where the $W_j(x)$ are fixed data-dependent unitaries (e.g., angle-embedding blocks), the H_j are Hermitian generators, and $\phi = (\phi_1, \dots, \phi_L) \in [-\pi, \pi]^L$ contains the L independent trainable encoding parameters. A circuit layer containing several trainable gates is decomposed into scalar-parameter factors of the form appearing in (2). The fixed unitary $W_0(x)$ allows the case $L = 0$ to represent an arbitrary fixed encoding. Data re-uploading circuits (Pérez-Salinas et al., 2020) with fixed input weights are obtained by repeating the same data-dependent block. Writing $g(z) = (z^\dagger E_i z)_{i=1}^k$, define the hypothesis and induced loss classes

$$\mathcal{G} = \{x \mapsto g(U\psi(x; \phi)) : U \in \mathbb{U}_Q, \phi \in [-\pi, \pi]^L\}, \quad (3)$$

$$\mathcal{F}_\gamma = \{(x, y) \mapsto \ell_\gamma(\mathcal{M}(h(x), y)) : h \in \mathcal{G}\}. \quad (4)$$

Theorem 2. *Consider an n -qubit QNN with unitary class $\mathbb{U}_Q$ and distance bound $b > 0$, POVMs $\{E_i\}_{i=1}^k$, and a trainable encoding of the form (2) with L parameters. For any $\gamma > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$ over an i.i.d. sample of size m , every hypothesis $h \in \mathcal{G}$ satisfies*

$$R(h) \leq \widehat{R}_\gamma(h) + \widetilde{O} \left(\frac{bE}{\gamma} \sqrt{\frac{n}{m}} + \sqrt{\frac{L}{m}} + \sqrt{\frac{\ln(2/\delta)}{m}} \right).$$

The first term is the margin-normalized capacity of the post-encoding QNN unitary class, as in Theorem 1, while the second term accounts for the trainable encoder. Setting $L = 0$ recovers the fixed-encoding case. The $\widetilde{O}$ notation suppresses logarithmic factors. Theorem 3 in Appendix A gives the explicit dependence.

Any variational classifier can be viewed as a trainable encoding followed by a fixed classifier, by absorbing all of its trainable gates into the encoding. This leads to the following corollary.

Corollary 1. *In the setting of Theorem 2 with $\mathbb{U}_Q = \{I\}$ and $EH\sqrt{L} \geq 2\gamma$, with probability at least $1 - \delta$ over an i.i.d. sample of size m , every hypothesis h satisfies*

$$R(h) \leq \widehat{R}_\gamma(h) + O \left(\sqrt{\frac{L}{m} \ln \frac{EH\sqrt{L}}{\gamma}} + \sqrt{\frac{\ln(2/\delta)}{m}} \right),$$

with absolute constants, where $H = (\sum_j \|H_j\|_\sigma^2)^{1/2}$. If every trainable gate is a Pauli rotation and the measurement is projective, the encoding term is $O(\sqrt{(L/m)\ln(L/\gamma)})$.

The complexity term is a parameter count and the margin enters only through $\widehat{R}_\gamma$ and the logarithm, as for a finite-dimensional Lipschitz class whose fat-shattering dimension at scale γ grows like $L \ln(1/\gamma)$ (Anthony and Bartlett, 2009), in contrast to the polynomial $1/\gamma$ of Theorem 2. It is a margin-loss analogue of the parameter-count bound of Caro et al. (2022). Explicit constants are given in Appendix A.8.

Proof sketch. The full proof is in Appendix A, and we outline the argument here. Write $X(\phi) = [\psi(x_1; \phi), \dots, \psi(x_m; \phi)] \in \mathbb{C}^{N \times m}$ for the encoded data matrix, whose columns are normalized states, so $\|X(\phi)\|_2 = \sqrt{m}$ for every ϕ . (1) Since $\ell_\gamma(\mathcal{M}(g(\cdot), y))$ is $(4E/\gamma)$ -Lipschitz uniformly in y (Hur and Park, 2025), covering numbers of $(\mathcal{F}_\gamma)_{|S}$ reduce to covering numbers of the set $\{UX(\phi) : U \in \mathbb{U}_Q, \phi \in [-\pi, \pi]^L\}$, which is no longer the image of a single data matrix. (2) The key step is a *two-scale covering* argument that restores the single-matrix structure.

Lemma 1 (Two-scale covering). *For any radii $r_1, r_2 > 0$,*

$$\mathcal{N}(\{UX(\phi)\}, r_1 + r_2, \|\cdot\|_2) \leq \mathcal{N}(\{X(\phi)\}, r_1, \|\cdot\|_2) \cdot \sup_{\phi'} \mathcal{N}(\{UX(\phi') : U \in \mathbb{U}_Q\}, r_2, \|\cdot\|_2).$$

The proof uses unitary invariance of the Frobenius norm. If $X(\phi)$ is close to a cover point $X(\phi')$, then $UX(\phi)$ is equally close to $UX(\phi')$, and the image $\mathbb{U}_Q X(\phi')$ can be covered by the original argument. (3) The unitary factor is bounded exactly as in the quantum-data case, since every $X(\phi')$ has Frobenius norm $\sqrt{m}$. (4) For the encoding factor, a telescoping argument shows the map $\phi \mapsto X(\phi)$ is $(\sqrt{m}H)$ -Lipschitz, so an r -net of the parameter cube $[-\pi, \pi]^L$ (of log-cardinality at most $L \ln(1 + \pi\sqrt{L}/r)$) induces a cover of $\{X(\phi)\}$. (5) Dudley’s entropy integral over the combined covering bound, with the discretization scale chosen proportional to $(bE/\gamma)\sqrt{\ln 4N^2}$, yields the stated complexity. Combining with (1) completes the proof.

The product form (2) is sufficient but not necessary: the proof only requires the encoder family to be uniformly Lipschitz in its trainable parameters. Fully classical neural encoders require a separate covering bound and are not covered by Theorem 2. We next evaluate margin- and parameter-based statistics across QNNs with varying encoder capacity.

5 Experimental Setup

We follow the predictor-evaluation methodology developed for classical deep networks (Jiang et al., 2019; Dziugaite et al., 2020). We train a large family of models spanning many hyperparameter axes, compute candidate complexity measures for each trained model, and score each measure by how well it predicts the generalization gap across controlled hyperparameter changes. Each hyperparameter is denoted in a bold font in the following subsections.

5.1 Datasets and Preprocessing

We construct three binary image-classification tasks: MNIST (LeCun et al., 2010) classes 1 vs. 6, KMNIST (Clanuwat et al., 2018) classes 0 vs. 5, and Fashion-MNIST (Xiao et al., 2017) classes 4 vs. 9. The **dataset** itself is treated as a hyperparameter in the evaluation below. For each class, 5,000 training and 500 test samples are drawn at random. Images are standardized, reduced to two dimensions by principal component analysis, and standardized again so that the two features can be used directly as rotation angles of quantum gates. The two-dimensional feature vector is tiled to match the register width. For example, $[x_0, x_1]$ becomes $[x_0, x_1, x_0, x_1, x_0, x_1]$ on six qubits.

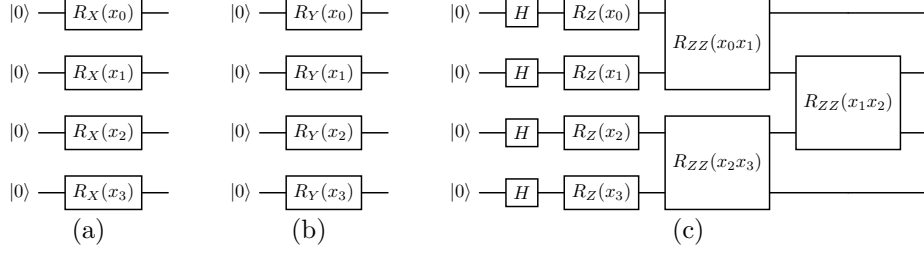


Figure 1: Four-qubit embedding circuits with input $\vec{x} = [x_0, x_1, x_2, x_3]$. (a) R_X angle embedding. (b) R_Y angle embedding. (c) One IQP embedding layer. One- and two-layer variants are used.

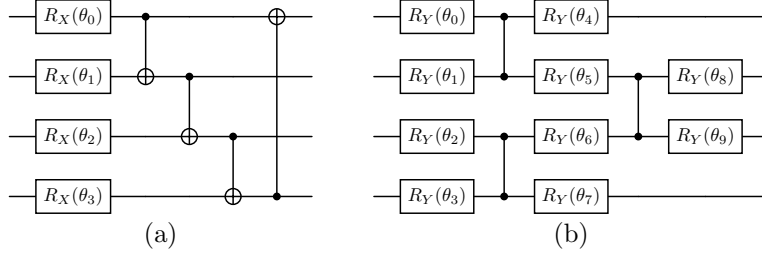


Figure 2: Four-qubit variational circuits with trainable parameters $\vec{\theta} = [\theta_0, \theta_1, \dots]$. (a) Basic entangler layer with R_X rotations. The R_Y variant substitutes R_Y gates, and strongly entangling layers substitute general $U3$ rotations in place of R_X rotations. (b) Simplified two-design layer.

5.2 Model Family

Each QNN is specified by four architectural hyperparameters.

Embedding circuit $U_E(x)$. Four fixed embeddings are used: R_X angle embedding, R_Y angle embedding, and IQP embeddings with one or two layers (Havlíček et al., 2019) (see Figure 1).

Variational circuit $V(\phi)$. Four ansätze are used: basic entangler layers with R_X or R_Y rotations, strongly entangling layers (Schuld et al., 2020), and simplified two-design layers (Cerezo et al., 2021) (see Figure 2).

Re-uploading depth and circuit width. We evaluate re-uploading depths $R \in \{4, 6, 8\}$ and circuit widths $n_q \in \{4, 6, 8\}$. The full circuit alternates embedding and variational blocks. The model prepares $(\prod_{i=0}^{R-1} V(\phi_i) U_E(x))|0\rangle^{\otimes n}$ and outputs the measurement distribution of qubit 0. If the outermost variational block is viewed as the post-encoding QNN unitary, the preceding trainable rotations play the role of the L encoding parameters in (2). The binary measurement on qubit 0 is projective, so $E = \sqrt{2}$ for every model.

5.3 Training

Three optimization hyperparameters are varied, namely the **batch size** $\in \{32, 64, 128\}$, the **learning rate** $\in \{0.001, 0.002, 0.005\}$, and the **random seed** $\in \{16, 13, 42, 65, 89\}$. Models are trained using Adam (Kingma and Ba, 2015) until the following convergence criterion is met. Let l_0 and σ_0 denote the mean and standard deviation of the most recent 500 mini-batch losses, respectively, and let l_1 denote the mean of the preceding 500 mini-batch losses. Training stops when $|l_0 - l_1| < \sigma_0/1000$. In total, the grid comprises $4 \times 4 \times 3 \times 3$ architectures, 3×3 optimization settings, 3 datasets, and 5 seeds, i.e., 19,440 training runs.

5.4 Complexity Measures

We compare five baseline complexity measures and two theorem-derived ones, all oriented so that *larger* values should indicate a *larger* generalization gap. Of the baselines, three contain margin information ($1/m_m$, $\sqrt{n_q}/m_m$, and n_p/m_m), while two are parameter-based (n_p and $n_p\sqrt{n_q}$). Here m_m is the median training margin, n_q is the number of qubits, and n_p is the total number of trainable parameters in the model. The factor $\sqrt{n_q}/m_m$ mirrors the explicit $\sqrt{n}/\gamma$ dependence in Theorem 2. The baselines isolate the ingredients of the bounds (margin, circuit width, and the parameter count that controls the encoding-capacity term), combining them only through simple products, so that comparisons among them attribute predictive power to the margin or to the parameter count. The median is used rather than the lower quartile to avoid infinite inverse margins for inexpressive models (Hur and Park, 2025). Results with lower-quantile margins are given in Appendix E.

The two theorem-derived measures combine the same ingredients in the functional forms that the bounds prescribe, and Section 6.2 compares them with the best baseline. The composite measure $\sqrt{n_q}/m_m + c\sqrt{n_p}$ matches the explicit dependence on γ , n_q and n_p in Theorem 2, with the margin and the encoding capacity as separate additive terms. The ratio c stands in for the unknown unitary radius and is chosen to maximize the generalization property prediction accuracy of the measure. The measure $n_p \ln(n_p/m_m)$ follows the Pauli-rotation form of Corollary 1 with $L = n_p$ and γ at the margin scale m_m , so that the margin enters only through a logarithm.

5.5 Evaluation Criteria

Conditional independence test. The conditional independence test of Jiang et al. (2019) measures the effect of a complexity measure μ on the generalization gap g while controlling for confounding by the hyperparameters. For a pair of hyperparameter settings (θ_1, θ_2) , define the sign random variables $V_\mu = \text{sign}(\mu(\theta_1) - \mu(\theta_2))$ and $V_g = \text{sign}(g(\theta_1) - g(\theta_2))$, and let U_S be uniformly distributed over the values of a subset S of hyperparameter types. The criterion is the worst-case normalized conditional mutual information over small conditioning sets,

$$\mathcal{K}(\mu) = \min_{U_S, |S| \leq 2} \hat{\mathcal{I}}(V_\mu, V_g \mid U_S), \quad \hat{\mathcal{I}} = \frac{\mathcal{I}(V_\mu, V_g \mid U_S)}{\mathcal{H}(V_g \mid U_S)}, \quad (5)$$

where $\mathcal{I}(\cdot, \cdot \mid \cdot)$ and $\mathcal{H}(\cdot \mid \cdot)$ denote conditional mutual information and conditional entropy, estimated from the empirical distribution over runs. Larger $\mathcal{K}(\mu)$ means μ carries more information about the generalization gap even after conditioning on the most adversarial pair of hyperparameters.

Sign-error distribution. Following Dziugaite et al. (2020), the sign-error quantifies how often the ordering of two models by μ disagrees with their ordering by generalization gap, across pairs of settings $(\theta_1, \theta_2) \in \mathcal{E}$ that differ in exactly one hyperparameter (random seeds excluded):

$$\widehat{SE} = \frac{1}{2} \mathbb{1}(n_{\text{eff}} \geq 12) \sum_{(\theta_1, \theta_2) \in \mathcal{E}} \bar{\kappa}(\theta_1, \theta_2) \times [1 - \text{sign}(g(\theta_1) - g(\theta_2)) \text{sign}(\mu(\theta_1) - \mu(\theta_2))]. \quad (6)$$

Each pair is weighted by $\bar{\kappa}(\theta_1, \theta_2)$, the normalization of $\kappa(\theta_1, \theta_2) = \max(0, \chi((\hat{g}(\theta_1) - \hat{g}(\theta_2))/2, m_{\text{te}}) - 0.5)$. The factor $\chi(\epsilon, m) = (1 - 2\exp(-2m\epsilon^2))^2$ lower-bounds the probability that the sign of the gap difference, estimated from a test set of size m , agrees with its sign under the true distribution (Dziugaite et al., 2020). Thus, only hyperparameter-setting pairs with sign-consistency probability at least 0.5 are retained. Some environments $\mathcal{E}$ may consequently contain too few reliable comparisons, so a sign-error is calculated only when the effective sample size $n_{\text{eff}} = (\sum_i \kappa_i)^2 / \sum_i \kappa_i^2$ is at least 12. Some complexity measures have a large fraction of environments where the change in a hyperparameter does not result in a change of complexity measure, thus the sign-error is evaluated as 0.5. These measures may have an unfair advantage due to the large number of environments having relatively low sign-error. To resolve such issues, we remove all environments where the complexity measure remains constant.

Complexity measure μ	$10^3 \mathcal{K}(\mu)$	$\widehat{SE}$, All		$\widehat{SE}$, Common	
		90%	Mean	90%	Mean
$1/m_m$	4.565	0.7124	0.4799	0.7225	0.4783
$\sqrt{n_q}/m_m$	3.543	0.7144	0.4801	0.7274	0.4787
n_p	0.1932	0.7724	0.5038	0.7724	0.5038
n_p/m_m	0.4683	0.7256	0.4879	0.7531	0.4962
$n_p\sqrt{n_q}$	0.1564	0.7724	0.5038	0.7724	0.5038

Table 1: Five baseline complexity measures: conditional independence test score (5) (higher is better) and 90th-percentile and mean sign-error (6) (lower is better) across all environments and across the Common environments in which the changed hyperparameter affects every measure (ansatz, number of re-uploadings, number of qubits). The maximum sign-error is 1 for every measure.

Computing $\widehat{SE}$ within each experimental environment (each fixed configuration of the remaining hyperparameters) yields a *distribution* of sign-errors per measure, whose upper tail reveals worst-case behavior. We report empirical cumulative distribution functions (ECDFs) and 90th-percentiles. 90th-percentiles indicate how well the complexity measure predicts the generalization gap for a large fraction of environments, thus indicating robustness as well as accuracy (Dziugaite et al. (2020)).

6 Results

Training diagnostics and the distribution of empirical generalization gaps over all runs are reported in Appendix B.

6.1 Baseline measures

Table 1 reports both evaluation criteria for the five baseline complexity measures. In the conditional independence test, the pure margin measure $1/m_m$ attains the highest score. It exceeds the score of the qubit-normalized margin measure $\sqrt{n_q}/m_m$ and every parameter-based measure. The mixed measure n_p/m_m underperforms compared to both inverse-margin measures, suggesting that multiplying margin information by a parameter count degrades rather than improves the signal.

The sign-error columns of Table 1 point the same way. No measure is completely robust (every measure exhibits environments with the maximal sign-error of 1), mirroring the classical findings of Dziugaite et al. (2020). At the 90th-percentile, however, the margin-based measures achieve markedly lower sign-error than the parameter-based measures.

Figure 3 decomposes the 90th-percentile sign-errors by the hyperparameter that is varied within each pair. Two observations stand out. First, the parameter-based measures n_p and $n_p\sqrt{n_q}$ are *constant* under changes of embedding, batch size, learning rate, and dataset, so those environments are excluded for them (Appendix C; dashes in Figure 3) and their scores rest on the three architecture axes, where the deficit is concentrated on the ansatz axis (0.878 versus 0.768 for $1/m_m$). Second, the margin-based measures show lower 90th-percentile values for optimization hyperparameters (batch size and learning rate) compared to other hyperparameters, which is desirable: a good complexity measure should respond to whatever affects generalization, whichever hyperparameter caused it (Dziugaite et al., 2020).

The overall advantage of margin-based measures could, in principle, stem solely from the axes on which parameter counts are constant. The Common columns of Table 1 therefore restrict the comparison to the hyperparameters that change *every* measure, namely the ansatz, the number of re-uploadings, and the number of qubits. The differences shrink, but the margin-based measures still dominate at the 90th-percentile and mean. Full empirical cumulative distributions of sign-errors are further detailed in Appendix H (Figure 6).

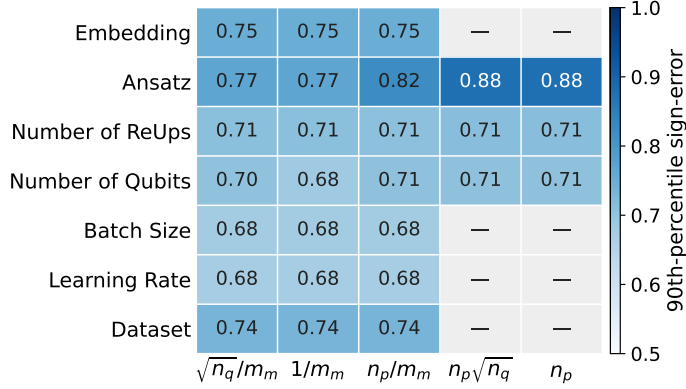


Figure 3: 90th-percentile values of the sign-error empirical distribution, decomposed by the hyperparameter varied within each pair. Lower is better.

Complexity measure μ	$10^3 \mathcal{K}(\mu)$	90% $\widehat{SE}$	
		All	Common
$1/m_m$	4.565	0.7124	0.7225
$\sqrt{n_q}/m_m + c^* \sqrt{n_p}$	3.316	0.7117	0.7212
$n_p \ln(n_p/m_m)$	0.3657	0.7321	0.7699

Table 2: Best baseline versus the two theorem-derived measures of Section 5.4: conditional independence score (5) (higher is better) and 90th-percentile sign-error over all and over the Common environments (lower is better); $c^* = 0.075$. Confidence intervals are given in Appendix G.

6.2 Theorem-derived measures

Table 2 compares the best baseline, $1/m_m$, with the two theorem-derived measures of Section 5.4, which combine the margin and the parameter count in the functional forms prescribed by the bounds. On the axes that leave n_p and n_q unchanged (embedding, batch size, learning rate, dataset) both order models exactly as $1/m_m$, so all differences arise on the architecture (Common) axes.

A composite complexity measure $\mu(c)$ of the form $\sqrt{n_q}/m_m + c\sqrt{n_p}$ that incorporates the number of qubits, number of parameters, and the margin following Theorem 2 is optimized in-sample through grid search to minimize the 90th-percentile sign-error. The minimum 90th-percentile sign-error of 0.7117 was achieved at $c^* = 0.075$, an improvement of 0.0007 over $1/m_m$ whose paired bootstrap interval includes zero $([-0.0026, 0.0012]$; Appendix G; held-out check in Appendix D). When applied to the conditional independence test, the optimized composite complexity measure scored $\mathcal{K}(\mu(c^*)) = 3.316 \times 10^{-3}$, which improves upon the parameter-based measures but not the pure margin. In fact, $\mathcal{K}$ decreases monotonically in c over the whole scanned range, from 3.55×10^{-3} at $c = 0$ to 6.6×10^{-5} at $c = 2$ (Table 3), so the composite never improves on its base $\sqrt{n_q}/m_m$ in this test. The conditional independence test is designed to expose this. It scores the information a measure carries about the gap once the hyperparameters are fixed, and since n_p is determined by three architecture hyperparameters, conditioning removes everything a term in n_p could contribute. What remains is a pull of the model-pair ordering toward that of n_p , one of the weakest measures in Table 1.

Since c depends on the unitary radius b , it should depend on the ansatz type and qubit number. Optimizing one weight c_{A,n_q} per (ansatz, n_q) pair lowers the 90th-percentile sign-error to 0.6988 on the five axes that leave c_{A,n_q} fixed (where $1/m_m$ scores 0.7052), but only to 0.7108 (versus 0.7124, within the bootstrap uncertainty) over all seven axes, with $\mathcal{K}(\mu(c_{A,n_q}^*)) = 2.630 \times 10^{-3}$, again below the pure margin measure (Appendix D).

Corollary 1 places the parameter count outside and the margin inside a logarithm. The measure $n_p \ln(n_p/m_m)$ is markedly worse than $1/m_m$ on both criteria (Table 2), because its ordering is dominated by

the parameter count: $\ln n_p$ ranges from 2.77 to 5.26 across the grid, whereas $\ln(1/m_m)$ has median 0.74, so the term $n_p \ln n_p$ exceeds the margin-dependent term $n_p \ln(1/m_m)$ in every run. Consequently $n_p \ln(n_p/m_m)$ orders 97.7% of the pairs of runs with different n_p exactly as n_p does (Kendall $\tau = 0.92$ with n_p , -0.06 with $1/m_m$) and inherits the conditional independence score of the parameter count.

Finally, we score the empirical margin loss $\hat{R}_\gamma$, i.e., the fraction of training points with margin at most γ . At each of $\gamma \in \{0.05, 0.1, 0.2\}$ it outperforms $1/m_m$ on both criteria, with $\mathcal{K}$ up to 1.56×10^{-2} and a 90th-percentile sign-error down to 0.706, a gain whose paired bootstrap interval excludes zero (Appendix F). The lower tail of the margin distribution thus carries more information about the gap than its median. Unlike $\hat{R}_\gamma$, however, $1/m_m$ does not depend on a margin level, which makes it the simpler summary of the margin distribution.

7 Conclusion

We extended margin-based generalization theory for quantum neural networks from quantum to classical data. For a fixed data encoding, the quantum-data margin bound of Hur and Park (2025) transfers without modification (Proposition 1). For trainable encodings, such as data re-uploading circuits, we proved a bound in which the margin-normalized term $(bE/\gamma)\sqrt{n/m}$ of the quantum-data case is accompanied by an additive encoding-capacity term of order $\sqrt{L/m}$ (Theorem 2), using a two-scale covering argument that treats the encoding parameters and the post-encoding unitary separately. When all trainable parameters lie in the encoding circuit, the complexity term reduces to a parameter count and the margin enters only logarithmically (Corollary 1).

Empirically, across 19,440 QNNs trained on binary tasks from MNIST, KMNIST, and Fashion-MNIST, the inverse margin is the strongest of the measures compared in predicting the generalization gap, under both the conditional independence test and the sign-error analysis. It outperforms the parameter-count baselines as well as the measures derived from Theorem 2 and Corollary 1, which combine the margin with the parameter count. The composite measure improves on the sign-error of the inverse margin only when its weight is tuned in-sample, and then by less than the bootstrap uncertainty. Since the parameter count is determined by the architecture hyperparameters on which the test conditions, it adds little information about the generalization gap beyond what the margin already carries. These findings extend the quantum-data observations of Hur and Park (2025) to classical data.

A natural follow-up question is how hardware noise affects these findings, as noise such as depolarizing noise contracts the measurement probabilities toward uniform and thereby shrinks the margin. On the theoretical side, Theorem 2 assumes that trainable gates and data-dependent gates are separate, as in Equation (2). Extending the analysis to encodings that process data and trainable parameters jointly, such as classical neural encoders (Hur et al., 2024) or rotation gates with trainable frequencies (Jaderberg et al., 2024), would bring a broader class of QNNs within the scope of the theory.

Code Availability

The code can be found in the following repository: <https://github.com/jyryu98/margin-gen-qml-classical-encoding>.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) ITRC (Information Technology Research Center) funded by the the Ministry of Science and ICT (MSIT), Korea (IITP-2026-RS-2026-25519864), by the IITP grant (No. 2019-0-00003, Research and Development of Core Technologies for Programming, Running, Implementing and Validating of Fault-Tolerant Quantum Computing System), by the National Research Foundation of Korea (RS-2025-02309510), by Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the

Ministry of Health & Welfare, Korea (RS-2025-25456722), and National Institute of Health (NIH) research project (2026-ER0901-00).

References

- Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., and Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409.
- Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.
- Banchi, L., Pereira, J., and Pirandola, S. (2021). Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2(4):040321.
- Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30.
- Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J., and Sweke, R. (2021). Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582.
- Caro, M. C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L., and Coles, P. J. (2022). Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1):4919.
- Cerezo, M., Sone, A., Volkoff, T., Cincio, L., and Coles, P. J. (2021). Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791.
- Choi, J., Hur, T., Jeong, S., Jung, K. L., Park, J. B., Lee, J., Jung, J. U., and Park, D. K. (2025). Optimizing quantum data embeddings for ligand-based virtual screening. *arXiv preprint arXiv:2512.16177*.
- Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. (2018). Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. (2020). In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733.
- Gil-Fuster, E., Eisert, J., and Bravo-Prieto, C. (2024). Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):2277.
- Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., and Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212.
- Hur, T., Araujo, I. F., and Park, D. K. (2024). Neural quantum embedding: Pushing the limits of quantum supervised learning. *Physical Review A*, 110(2):022411.
- Hur, T. and Park, D. K. (2025). Understanding generalization in quantum machine learning with margins. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR.
- Jaderberg, B., Gentile, A. A., Berrada, Y. A., Shishenina, E., and Elfving, V. E. (2024). Let quantum neural networks choose their own frequencies. *Physical Review A*, 109(4):042421.
- Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.
- Kim, Y., Im, C., Kim, T., Hur, T., and Park, D. K. (2026). Multi-channel convolutional neural quantum embedding. *Advanced Quantum Technologies*, 9(1):e00575.

- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
- LeCun, Y., Cortes, C., and Burges, C. J. (2010). MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist>.
- Liu, H., Hur, T., Zhang, S., Che, L., Long, X., Wang, X., Huang, K., Fan, Y.-a., Zheng, Y., Feng, Y., et al. (2025). Neural quantum embedding via deterministic quantum computation with one qubit. *Physical Review Letters*, 135(8):080603.
- Lloyd, S., Schuld, M., Ijaz, A., Izaac, J., and Killoran, N. (2020). Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT Press, 2nd edition.
- Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*.
- Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., and Latorre, J. I. (2020). Data re-uploading for a universal quantum classifier. *Quantum*, 4:226.
- Rodriguez-Grasa, P., Caro, M. C., Eisert, J., Gil-Fuster, E., Schreiber, F. J., and Bravo-Prieto, C. (2026). A pac-bayesian approach to generalization for quantum models. *arXiv preprint arXiv:2603.22964*.
- Schuld, M., Bocharov, A., Svore, K. M., and Wiebe, N. (2020). Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308.
- Schuld, M., Sweke, R., and Meyer, J. J. (2021). Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430.
- Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.
- Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550.

A Proof of Theorem 2

Throughout, n is the number of qubits, $N = 2^n$, m the sample size, and k the number of classes. Here $E = (\sum_{i=1}^k \|E_i\|_\sigma^2)^{1/2}$ and $b > 0$ satisfies $\|U - U_{\text{ref}}\|_{2,1} \leq b$ for all $U \in \mathbb{U}_Q$. We write $X(\phi) = [\psi(x_1; \phi), \dots, \psi(x_m; \phi)] \in \mathbb{C}^{N \times m}$ for the encoded sample matrix. Since its columns are normalized quantum states, $\|X(\phi)\|_2 = \sqrt{m}$ for every ϕ , where $\|\cdot\|_2$ on matrices denotes the Frobenius norm. For a matrix X , we write $\mathbb{U}_Q X = \{UX : U \in \mathbb{U}_Q\}$ for its image under the unitary class.

The argument first establishes the following explicit-logarithmic version of Theorem 2 under two side conditions. Appendix A.7 then removes those conditions, which only exclude regimes where Theorem 2 holds for elementary reasons.

Theorem 3 (Explicit-logarithmic version of Theorem 2). *In the setting of Theorem 2, assume $bE \geq \gamma/4$ and $m \geq m_0 = 144b^2E^2 \ln(4N^2)/\gamma^2$. Then the empirical Rademacher complexity in (1) obeys*

$$\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) \in O\left(\frac{bE}{\gamma} \sqrt{\frac{n}{m}} \left(1 + \ln \frac{\gamma\sqrt{m}}{bE\sqrt{n}}\right) + \sqrt{\frac{L}{m}} \left(1 + \ln \left(1 + \frac{8\pi EH\sqrt{L}}{\gamma}\right)\right)\right).$$

A.1 Reduction to Matrix Covering

The quantum-data analysis of Hur and Park (2025) rests on two facts. First, for any label y , the map $z \mapsto \ell_\gamma(\mathcal{M}(g(z), y))$ is $(4E/\gamma)$ -Lipschitz in $\|\cdot\|_2$ on normalized states, since ℓ_γ is $(1/\gamma)$ -Lipschitz, $\mathcal{M}$ is 2-Lipschitz in ℓ_∞ , and each measurement function $g_i(z) = z^\dagger E_i z$ is $2\|E_i\|_\sigma$ -Lipschitz. Second, the image $\mathbb{U}_Q X$ of a fixed data matrix admits the covering bound (their Equation (9), obtained from Maurey sparsification (Zhang, 2002) applied to complex matrices)

$$\ln \mathcal{N}(\{UX : U \in \mathbb{U}_Q\}, r, \|\cdot\|_2) \leq \left\lceil \frac{2mb^2}{r^2} \right\rceil \ln 4N^2 \quad (7)$$

for any matrix X with $\|X\|_2 = \sqrt{m}$ and any $r > 0$.

For the class (3), the Lipschitz property gives, for any $\epsilon > 0$,

$$\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2) \leq \ln \mathcal{N}(\{UX(\phi) : U \in \mathbb{U}_Q, \phi \in [-\pi, \pi]^L\}, \frac{\epsilon\gamma}{4E}, \|\cdot\|_2), \quad (8)$$

but the right-hand side is no longer the image of a single matrix. The data matrix moves with ϕ .

A.2 Two-Scale Covering

Lemma 2 (Lemma 1, restated). *For any $r_1, r_2 > 0$,*

$$\mathcal{N}(\{UX(\phi)\}, r_1 + r_2, \|\cdot\|_2) \leq \mathcal{N}(\{X(\phi) : \phi \in [-\pi, \pi]^L\}, r_1, \|\cdot\|_2) \cdot \sup_{\phi'} \mathcal{N}(\{UX(\phi') : U \in \mathbb{U}_Q\}, r_2, \|\cdot\|_2).$$

Here the first covering number is over covers by points of $\{X(\phi)\}$ itself (internal covers): the grid of Lemma 4 is one, and an internal covering number is at most the external one at half the radius.

Proof. Take $C_\Phi \subset [-\pi, \pi]^L$ such that $\{X(\phi') : \phi' \in C_\Phi\}$ is an r_1 -cover of $\{X(\phi)\}$. The grid construction of Lemma 4 below produces the cover from points *inside* the cube, so every cover point is itself an admissible data matrix. For each $\phi' \in C_\Phi$, let $C_{\phi'}$ be an r_2 -cover of the image $\mathbb{U}_Q X(\phi')$. Given any (U, ϕ) , choose $\phi' \in C_\Phi$ with $\|X(\phi) - X(\phi')\|_2 \leq r_1$. By unitary invariance of the Frobenius norm, $\|UX(\phi) - UX(\phi')\|_2 = \|X(\phi) - X(\phi')\|_2 \leq r_1$. Then choose $Z \in C_{\phi'}$ with $\|UX(\phi') - Z\|_2 \leq r_2$. Hence $\bigcup_{\phi' \in C_\Phi} C_{\phi'}$ is an $(r_1 + r_2)$ -cover of cardinality at most the claimed product. $\square$

Splitting the radius in (8) as $\frac{\epsilon\gamma}{4E} = \frac{\epsilon\gamma}{8E} + \frac{\epsilon\gamma}{8E}$ and applying the lemma,

$$\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2) \leq \ln \mathcal{N}\left(\{X(\phi)\}, \frac{\epsilon\gamma}{8E}, \|\cdot\|_2\right) + \sup_{\phi'} \ln \mathcal{N}\left(\{UX(\phi') : U \in \mathbb{U}_Q\}, \frac{\epsilon\gamma}{8E}, \|\cdot\|_2\right). \quad (9)$$

Since every $X(\phi')$ satisfies $\|X(\phi')\|_2 = \sqrt{m}$, the second term follows from (7) at radius $\epsilon\gamma/8E$:

$$\sup_{\phi'} \ln \mathcal{N}\left(\{UX(\phi')\}, \frac{\epsilon\gamma}{8E}, \|\cdot\|_2\right) \leq \left\lceil \frac{128 mb^2 E^2}{\epsilon^2 \gamma^2} \right\rceil \ln 4N^2. \quad (10)$$

A.3 Lipschitz Property of the Encoding

Lemma 3. *For all $\phi, \phi' \in [-\pi, \pi]^L$,*

$$\|\psi(x; \phi) - \psi(x; \phi')\|_2 \leq H \|\phi - \phi'\|_2,$$

with $H = \left(\sum_{j=1}^L \|H_j\|_\sigma^2\right)^{1/2}$, and consequently $\|X(\phi) - X(\phi')\|_2 \leq \sqrt{m} H \|\phi - \phi'\|_2$.

Proof. For unitaries $A_1, \dots, A_L$ and $B_1, \dots, B_L$, a telescoping argument gives $\left\|\prod_j A_j - \prod_j B_j\right\|_\sigma \leq \sum_j \|A_j - B_j\|_\sigma$, since all remaining factors have unit spectral norm. The fixed interleaved unitaries $W_j(x)$ cancel in each telescoping step. For a single layer,

$$\|e^{iHa} - e^{iHb}\|_\sigma = \left\|iH \int_b^a e^{iHt} dt\right\|_\sigma \leq \|H\|_\sigma |a - b|.$$

Combining the two facts and applying the Cauchy-Schwarz inequality,

$$\|\psi(x; \phi) - \psi(x; \phi')\|_2 \leq \sum_{j=1}^L \|H_j\|_\sigma |\phi_j - \phi'_j| \leq H \|\phi - \phi'\|_2.$$

The matrix statement follows because each of the m columns changes by at most $H \|\phi - \phi'\|_2$ in Euclidean norm. $\square$

More general parameterized layers. The subsequent argument uses only the uniform Lipschitz property in Lemma 3. It therefore also applies to encoders $V(x; \phi) = \prod_{j=1}^L V_j(x; \phi_j)$ satisfying $\|\partial_{\phi_j} V_j(x; \phi_j)\|_\sigma \leq c_j$ uniformly in x and ϕ_j , with H replaced by $(\sum_j c_j^2)^{1/2}$. This includes data-parameter coupled rotations such as $e^{i\phi_j x_\ell H}$ when the data are bounded. Encoders implemented by a separate classical neural network require a covering bound for that network and are outside Theorem 2.

A.4 Covering the Parameter Cube

Lemma 4. *For any $r > 0$,*

$$\ln \mathcal{N}([-\pi, \pi]^L, r, \|\cdot\|_2) \leq L \ln \left(1 + \frac{\pi\sqrt{L}}{r}\right).$$

Proof. An axis-aligned grid of spacing $2r/\sqrt{L}$ inside $[-\pi, \pi]^L$ matches each coordinate to within $r/\sqrt{L}$ and is therefore an r -cover in $\|\cdot\|_2$, with $\lceil \pi\sqrt{L}/r \rceil$ points per axis, giving $\ln \mathcal{N} \leq L \ln \lceil \pi\sqrt{L}/r \rceil \leq L \ln(1 + \pi\sqrt{L}/r)$. $\square$

We state the bound in the $\ln(1 + \cdot)$ form, standard for covering numbers of bounded sets (compare $\mathcal{N}(B_2^d, r, \|\cdot\|_2) \leq (1 + 2R/r)^d$ for the radius- R ball). The covering number of a nonempty set is at least one, so a valid upper bound on $\ln \mathcal{N}$ must remain nonnegative for every $r > 0$. An expression of the form $L \ln(K/\epsilon)$ fails this for $\epsilon > K$, while $\ln(1 + \cdot)$ needs no side condition. In the regime $\epsilon \ll K$ relevant inside Dudley's integral, the two forms agree up to constants.

By Lemma 3, an $\frac{\epsilon\gamma}{8EH\sqrt{m}}$ -cover of the cube induces an $\frac{\epsilon\gamma}{8E}$ -cover of $\{X(\phi)\}$, so

$$\ln \mathcal{N}\left(\{X(\phi)\}, \frac{\epsilon\gamma}{8E}, \|\cdot\|_2\right) \leq L \ln\left(1 + \frac{K}{\epsilon}\right), \quad (11)$$

where $K = 8\pi EH\sqrt{mL}/\gamma$.

A.5 Covering Bound for the Hypothesis Class

Under the assumption $bE \geq \gamma/4$ of Theorem 3, for all $0 < \epsilon \leq \sqrt{m}$ one has $128mb^2E^2/(\epsilon^2\gamma^2) \geq 8$, so $\lceil x \rceil \leq \frac{9}{8}x$ (valid for $x \geq 8$) absorbs the ceiling in (10). Radii $\epsilon > \sqrt{m}$ are irrelevant since $(\mathcal{F}_\gamma)_{|S} \subseteq [0, 1]^m$ has diameter at most $\sqrt{m}$. Combining (9), (10), and (11) gives, for all $0 < \epsilon \leq \sqrt{m}$,

$$\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2) \leq L \ln\left(1 + \frac{K}{\epsilon}\right) + \frac{A}{\epsilon^2}, \quad (12)$$

with $A = 144mb^2E^2 \ln(4N^2)/\gamma^2$.

A.6 Dudley's Entropy Integral

Dudley's entropy integral (Bartlett et al., 2017, Lemma A.5) gives

$$\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_{\alpha}^{\sqrt{m}} \sqrt{\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \beta, \|\cdot\|_2)} d\beta \right).$$

The lemma is stated for sets containing the origin, and applying it to $(\mathcal{F}_\gamma)_{|S} \cup \{0\}$ increases each covering number by at most one element, affecting only absolute constants. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ on (12), the two contributions are

$$\frac{12}{m} \int_{\alpha}^{\sqrt{m}} \frac{\sqrt{A}}{\beta} d\beta = \frac{12\sqrt{A}}{m} \ln \frac{\sqrt{m}}{\alpha} \quad \text{and} \quad \frac{12}{m} \int_{\alpha}^{\sqrt{m}} \sqrt{L \ln(1 + K/\beta)} d\beta \leq 12\sqrt{\frac{L}{m}} \left(\sqrt{\ln(1 + K')} + 0.89 \right),$$

where $K' = K/\sqrt{m} = 8\pi EH\sqrt{L}/\gamma$. The second bound follows by extending the integral down to 0, which is allowed because its integrand is nonnegative and integrable at 0: substituting $\beta = \sqrt{m}u$ and using $\ln(1 + K'/u) \leq \ln(1 + K') + \ln(1/u)$ for $0 < u \leq 1$ together with $\int_0^1 \sqrt{\ln(1/u)} du = \sqrt{\pi}/2 \leq 0.89$, exactly as in Appendix A.8. The cutoff α is thus needed only for the unitary term. Choose

$$\alpha = \sqrt{\frac{A}{m}} = \frac{12bE\sqrt{\ln 4N^2}}{\gamma},$$

which is admissible ($\alpha \leq \sqrt{m}$) precisely when $m \geq m_0$, the sample-size condition of Theorem 3. Then

$$\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) \leq \frac{4\sqrt{A}}{m} + \frac{12\sqrt{A}}{m} \ln \frac{m}{\sqrt{A}} + 12\sqrt{\frac{L}{m}} \left(\sqrt{\ln(1 + K')} + 0.89 \right), \quad (13)$$

with

$$\frac{\sqrt{A}}{m} = \frac{12bE}{\gamma} \sqrt{\frac{\ln 4N^2}{m}}.$$

Since $\ln 4N^2 = 2(n+1)\ln 2 = \Theta(n)$ and $\sqrt{a} + 0.89 \leq \sqrt{2(1+a)}$ for $a \geq 0$,

$$\begin{aligned} \Re((\mathcal{F}_\gamma)_{|S}) &\in O\left(\frac{bE}{\gamma}\sqrt{\frac{n}{m}}\left(1 + \ln\frac{\gamma\sqrt{m}}{bE\sqrt{n}}\right) + \sqrt{\frac{L}{m}}\left(1 + \ln\left(1 + \frac{8\pi EH\sqrt{L}}{\gamma}\right)\right)\right) \\ &\in \tilde{O}\left(\frac{bE}{\gamma}\sqrt{\frac{n}{m}} + \sqrt{\frac{L}{m}}\right). \end{aligned}$$

This establishes Theorem 3, and substituting into the margin bound (1) yields Theorem 2 whenever $bE \geq \gamma/4$ and $m \geq m_0$. $\square$

A.7 Removing the Side Conditions

It remains to prove Theorem 2 when either side condition of Theorem 3 fails.

Small samples. Suppose $m < m_0$. Then $(bE/\gamma)\sqrt{\ln(4N^2)/m} > 1/12$, and since $\ln 4N^2 = 2(n+1)\ln 2 \leq 4n\ln 2$ for $n \geq 1$, the first term of Theorem 2 satisfies $(bE/\gamma)\sqrt{n/m} > 1/20$. Once the constant hidden in the $\tilde{O}$ is at least 20, the right-hand side of the bound exceeds one, and the bound holds trivially because $R(h) \leq 1$.

Weak unitary classes. Suppose $0 < bE < \gamma/4$ and $m \geq m_0$. The ceiling in (10) can then no longer be absorbed multiplicatively, so we keep it via $\lceil x \rceil \leq x+1$, and (12) acquires the additive term $\ln 4N^2$. This term is active only at small radii. Since $\|U - U_{\text{ref}}\|_\sigma \leq \|U - U_{\text{ref}}\|_2 \leq \|U - U_{\text{ref}}\|_{2,1} \leq b$, every matrix in the image $\mathbb{U}_Q X(\phi')$ satisfies $\|UX(\phi') - U_{\text{ref}}X(\phi')\|_2 \leq \|U - U_{\text{ref}}\|_\sigma \|X(\phi')\|_2 \leq b\sqrt{m}$, so the single point $U_{\text{ref}}X(\phi')$ covers $\mathbb{U}_Q X(\phi')$ at radius $b\sqrt{m}$. The unitary factor in (9) is therefore zero at all radii $\epsilon\gamma/(8E) \geq b\sqrt{m}$, that is for all $\epsilon \geq D = 8bE\sqrt{m}/\gamma$. The additive term hence contributes to Dudley's entropy integral at most

$$\frac{12}{m} \int_0^D \sqrt{\ln 4N^2} d\beta = \frac{96bE}{\gamma} \sqrt{\frac{\ln 4N^2}{m}} \in O\left(\frac{bE}{\gamma} \sqrt{\frac{n}{m}}\right),$$

which is absorbed into the first term of Theorem 2. The remainder of the proof proceeds unchanged, completing the proof for every $b > 0$. When $b = 0$, the post-encoding unitary class is a singleton; applying Dudley's integral directly to the encoder cover (11) gives the simpler encoder-only scaling $\tilde{O}(\sqrt{L/m})$. $\square$

A.8 Proof of Corollary 1

With $\mathbb{U}_Q = \{I\}$, the reduction (8) reads $\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2) \leq \ln \mathcal{N}(\{X(\phi) : \phi \in [-\pi, \pi]^L\}, \epsilon\gamma/4E, \|\cdot\|_2)$, with no unitary factor. By Lemma 3 an r -cover of the cube with $r = \epsilon\gamma/(4EH\sqrt{m})$ induces an $\epsilon\gamma/4E$ -cover of $\{X(\phi)\}$, so Lemma 4 gives

$$\ln \mathcal{N}((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2) \leq L \ln \left(1 + \frac{K_0}{\epsilon}\right), \quad K_0 = \frac{4\pi EH\sqrt{mL}}{\gamma}.$$

Since $\beta \mapsto \sqrt{\ln(1 + K_0/\beta)}$ is integrable at 0, Dudley's integral may be taken with lower limit $\alpha \rightarrow 0$, which removes the $4\alpha/\sqrt{m}$ term. Applying it to $(\mathcal{F}_\gamma)_{|S} \cup \{0\}$ and substituting $\beta = \sqrt{m}u$, with $K'_0 = K_0/\sqrt{m} = 4\pi EH\sqrt{L}/\gamma$ and the elementary bound $\ln(1 + K'_0/u) \leq \ln(1 + K'_0) + \ln(1/u)$ for $0 < u \leq 1$,

$$\begin{aligned} \Re((\mathcal{F}_\gamma)_{|S}) &\leq \frac{12}{\sqrt{m}} \int_0^1 \left(\sqrt{L \ln(1 + K'_0/u)} + \sqrt{\ln 2}\right) du \\ &\leq 12\sqrt{\frac{L}{m}} \left(\sqrt{\ln(1 + K'_0)} + \frac{\sqrt{\pi}}{2}\right) + \frac{12\sqrt{\ln 2}}{\sqrt{m}}, \end{aligned}$$

using $\int_0^1 \sqrt{\ln(1/u)} du = \sqrt{\pi}/2 \leq 0.89$. Substituting into the margin bound (1) gives, with probability at least $1 - \delta$,

$$R(h) \leq \hat{R}_\gamma(h) + 24\sqrt{\frac{L}{m}} \left(\sqrt{\ln\left(1 + \frac{4\pi EH\sqrt{L}}{\gamma}\right)} + 0.89 \right) + \frac{20}{\sqrt{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \quad (14)$$

When $EH\sqrt{L} \geq 2\gamma$ the logarithm is at least $\ln(1 + 8\pi) > 3$, so the bracket is at most a constant multiple of $\sqrt{\ln(1 + K'_0)}$ and $\ln(1 + 4\pi x) \leq 5 \ln x$ for $x \geq 2$; this yields Corollary 1 with absolute constants. If every trainable gate is a Pauli rotation $\exp(-i\theta\sigma/2)$, then $\|H_j\|_\sigma = 1/2$, $H = \sqrt{L}/2$, and $K'_0 = 2\pi EL/\gamma$. If, in addition, the measurement is projective with k outcomes, each E_i is an orthogonal projector, so $\|E_i\|_\sigma \leq 1$, $E \leq \sqrt{k}$, and $K'_0 \leq 2\pi\sqrt{k}L/\gamma$. We may take $\gamma < 1$, since margins never exceed 1 and $\hat{R}_\gamma(h) = 1$ otherwise; then $L/\gamma \geq 2$ for $L \geq 2$, and $\ln(1 + K'_0) \leq \ln(1 + 2\pi\sqrt{k}) + \ln(L/\gamma) \leq c_k \ln(L/\gamma)$ with $c_k = 1 + \ln(1 + 2\pi\sqrt{k})/\ln 2$ ($c_2 < 4.4$ for binary classification). The same lower bound $\ln(L/\gamma) \geq \ln 2$ absorbs the additive constant 0.89 in (14), giving the stated $O(\sqrt{(L/m)\ln(L/\gamma)})$ form, with a constant depending only on k . $\square$

B Training Diagnostics

Figure 4 summarizes the training results in all runs. Most runs achieve high training and test accuracy, with a population of poorly performing outliers concentrated among the smallest architectures. For two of the three datasets, the empirical generalization gap (training minus test accuracy) is distributed approximately symmetrically around zero rather than being predominantly positive. The gaps in this regime are small, so fluctuations from the finite test set can be comparable to the gaps themselves. The κ -weighting in Equation (6) discounts comparisons for which the ordering is not reliably determined.

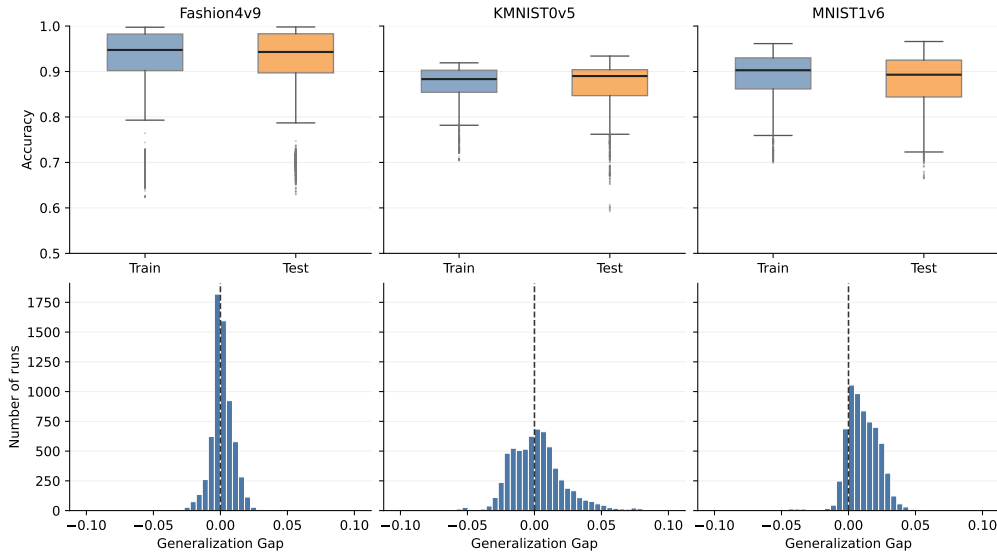


Figure 4: Accuracies and generalization gaps of all 19,440 runs. The training accuracy and test accuracies are represented as boxplots. The box extends from the first quartile to the third, with a line at the median. Error bars represent $1.5\times$ the inter-quartile range added to the box boundaries.

C Additional Experimental Details

Evaluation details. For the conditional independence test defined in Equation (5), the empirical joint distribution over runs is used to estimate the conditional mutual information and conditional entropy, following

Hur and Park (2025). For the sign-error defined in Equation (6), the function $\chi(\cdot, m_{te})$ is the sign-consistency probability bound of Dziugaite et al. (2020) evaluated with the test-set size $m_{te} = 1000$. Pairs are formed only between settings differing in exactly one hyperparameter, seeds are aggregated within each setting, and environments with effective sample size $n_{eff} < 12$ are discarded. Environments where the complexity measure is constant about the change in the hyperparameter are also discarded. In Equation (5), S ranges over all eight hyperparameter types of Section 5, including the random seed, i.e. 36 conditioning sets (8 single types and 28 pairs); for every measure reported, the minimizing set does not involve the seed, so the 28 seed-free sets give the same $\mathcal{K}$. Of the 30,732 environments that contain at least one retained pair ($\kappa > 0$), 24,909 satisfy $n_{eff} \geq 12$ (median $n_{eff} = 20.4$ among them, of at most 25). The margin-based measures are scored on these 24,909 environments (11,104 Common), whereas n_p and $n_p\sqrt{n_q}$ are scored on 10,443 in both columns, because n_p is also constant between the two basic-entangler ansätze, which removes 661 further environments. Both criteria depend only on how a measure orders pairs of models, so strictly increasing transformations of a measure, such as the square root in Corollary 1 or in $\sqrt{n_p}$, leave all scores unchanged.

Margin computation. Margins are computed on the training set after convergence as $\mathcal{M}(h(x), y) = h(x)_y - \max_{j \neq y} h(x)_j$, where $h(x)$ is the measurement distribution of qubit 0, and m_m denotes the median of this distribution.

Software and hardware. Simulations were performed with the `default.qubit` statevector simulator in PennyLane 0.42.3. PyTorch 2.8.0 was the main machine learning framework used. Experiments were carried out on a desktop computer equipped with an AMD Ryzen 7 7800X3D 8-core processor and 32GBs of RAM.

D Composite Measure Optimization

We optimize $\mu(c) = \sqrt{n_q}/m_m + c\sqrt{n_p}$ to minimize the 90th-percentile sign-error by grid-searching c over $[0.01, 0.1]$ with increments of 0.005. The sign-errors formed a U-shape across this sweep as shown in Figure 5(a), achieving a minimum 90th-percentile sign-error of 0.7117 at $c^* = 0.075$.

Table 3 reports both criteria along a wider sweep of c : the 90th-percentile sign-error is U-shaped with its minimum at $c^* = 0.075$, whereas $\mathcal{K}$ decreases monotonically over the whole sweep, from its $c = 0$ value $\mathcal{K}(\sqrt{n_q}/m_m)$ to 6.6×10^{-5} at $c = 2$, below even the parameter-count ordering recovered as $c \rightarrow \infty$, so no choice of c improves the conditional independence score.

c	0	0.01	0.025	0.05	0.075	0.1	0.15	0.2	0.3	0.5	1	2	∞
$10^3 \mathcal{K}(\mu(c))$	3.546	3.518	3.485	3.407	3.316	3.269	3.139	3.006	2.731	2.284	1.130	0.066	0.194
$\widehat{SE}_{90}$, All	0.7144	0.7136	0.7134	0.7121	0.7117	0.7127	0.7129	0.7138	0.7154	0.7183	0.7233	0.7294	0.7724
$\widehat{SE}_{90}$, Com.	0.7274	0.7250	0.7247	0.7221	0.7212	0.7233	0.7237	0.7257	0.7310	0.7385	0.7486	0.7592	0.7724

Table 3: Composite measure $\mu(c) = \sqrt{n_q}/m_m + c\sqrt{n_p}$ along a sweep of c : conditional independence score (higher is better) and 90th-percentile sign-error over all and over the Common environments (lower is better). $c = 0$ is $\sqrt{n_q}/m_m$; the column $c = \infty$ is the ordering of $\sqrt{n_p}$, i.e., of n_p ; its sign-errors are over the 10,443 environments in which n_p varies (finite c : all 24,909). For reference, $\mathcal{K}(1/m_m) = 4.570 \times 10^{-3}$ with sign-errors 0.7124 and 0.7225.

We also optimize $\mu(c_{A,n_q}) = \sqrt{n_q}/m_m + c_{A,n_q}\sqrt{n_p}$, where each (qubit number, ansatz) pair has its own independent c_{A,n_q} . Each setting had its c value grid-searched across $[0, 0.2]$ with increments of 0.005. The 90th-percentile sign-errors for each setting pair for this sweep is shown in Figure 5(b). Using the optimum c_{A,n_q} values, the aggregate sign-error 90th-percentile is 0.6988. This aggregate runs over the environments in which c_{A,n_q} is constant, i.e., those varying embedding, re-uploadings, batch size, learning rate, or dataset (17,199 environments), not over the seven axes of Tables 1 and 2; on the same five axes $1/m_m$ scores 0.7052 and the single- c composite 0.7012, while over all seven axes $\mu(c_{A,n_q})$ scores 0.7108 versus 0.7124 for $1/m_m$ (paired bootstrap difference -0.0017 , 95% interval $[-0.0036, 0.0001]$). The gain is confined to the

measure	$\mathcal{K}$ p50	$\mathcal{K}$ p25	$\mathcal{K}$ p10	$\widehat{SE}$ p50	$\widehat{SE}$ p25	$\widehat{SE}$ p10
$1/m_m$	4.57e-3	1.10e-2	1.09e-2	0.7124	0.7017	0.7082
$\sqrt{n_q}/m_m$	3.55e-3	9.00e-3	1.05e-2	0.7144	0.7044	0.7092
n_p/m_m	4.69e-4	5.86e-3	9.79e-3	0.7256	0.7118	0.7071
$n_p\sqrt{n_q}$	1.57e-4	1.57e-4	1.57e-4	0.7724	0.7724	0.7724
n_p	1.94e-4	1.94e-4	1.94e-4	0.7724	0.7724	0.7724
$n_p \ln(n_p/m_m)$	3.66e-4	3.38e-3	5.76e-3	0.7321	0.7267	0.7237
composite ($c=0.075$)	3.32e-3	8.89e-3	1.06e-2	0.7117	0.7028	0.7079

Table 4: Margin-statistic ablation: $\mathcal{K}$ (higher is better) and 90% $\widehat{SE}$ over all environments (lower is better) with the margin taken at the 50th, 25th, and 10th percentile. Runs with $m_m=0$: 0 (p50), 1,149 (p25), 9,279 (p10). The parameter-only measures n_p , $n_p\sqrt{n_q}$ do not use the margin, so their scores are constant across statistics.

F Empirical Margin Loss

Table 5 evaluates the empirical margin loss $\widehat{R}_\gamma(h) = m^{-1} \sum_i \mathbb{1}(\mathcal{M}(h(x_i), y_i) \leq \gamma)$ as a complexity measure at three fixed margin levels, with $1/m_m$ for reference. Because negative sample margins are recorded as zero, the count $\#\{\mathcal{M} \leq \gamma\}$ is unchanged by that convention for every $\gamma > 0$, so $\widehat{R}_\gamma$ is the exact empirical margin loss and is finite for every run. It exceeds the inverse median margin on both criteria at all three levels (paired bootstrap of the 90th-percentile difference at $\gamma = 0.1$: -0.0069 , 95% interval $[-0.0111, -0.0034]$; Appendix G); brackets are 95% bootstrap intervals over resampled environments (an independent bootstrap draw from that of Table 6, so the $1/m_m$ interval differs in the last digit).

measure	$\mathcal{K}(\mu)$	90% $\widehat{SE}$ All [95% CI]	90% $\widehat{SE}$ Common
$\widehat{R}_{\gamma=0.05}$	1.56×10^{-2}	0.7059 [0.7011, 0.7100]	0.7082
$\widehat{R}_{\gamma=0.1}$	1.34×10^{-2}	0.7056 [0.7011, 0.7090]	0.7113
$\widehat{R}_{\gamma=0.2}$	9.63×10^{-3}	0.7082 [0.7042, 0.7122]	0.7174
$1/m_m$	4.57×10^{-3}	0.7124 [0.7080, 0.7166]	0.7225

Table 5: Empirical margin loss $\widehat{R}_\gamma$ as a predictor. Higher $\mathcal{K}$ and lower $\widehat{SE}$ are better.

G Bootstrap Confidence Intervals

Table 6 reports 95% bootstrap intervals for the 90th-percentile sign-error, obtained by resampling the sign-error environments with replacement (2,000 resamples). The intervals separate the pure-margin measures ($1/m_m$, $\sqrt{n_q}/m_m$, and the composite; upper limit ≤ 0.72) from the parameter-count measures (n_p , $n_p\sqrt{n_q}$; lower limit ≥ 0.76), with n_p/m_m and $n_p \ln(n_p/m_m)$ falling in between, while the differences among the pure-margin measures—including the 0.0007 edge of the composite over $1/m_m$ —lie within the intervals. Because the measures are scored on the same environments, the sharper test is a paired bootstrap of the difference in 90th-percentile sign-error on the shared environments (24,909; 24,897 for $\widehat{R}_{\gamma=0.1}$; 2,000 resamples): composite minus $1/m_m$ is -0.0007 with 95% interval $[-0.0026, 0.0012]$, and $\widehat{R}_{\gamma=0.1}$ minus $1/m_m$ is -0.0069 $[-0.0111, -0.0034]$ (-0.0113 $[-0.0175, -0.0072]$ on the Common environments). Resampling the (varied hyperparameter, base configuration) clusters instead of individual environments leaves these intervals essentially unchanged (composite $[-0.0028, 0.0012]$, $\widehat{R}_{\gamma=0.1}$ $[-0.0115, -0.0028]$). Confidence intervals for the conditional independence score $\mathcal{K}$, obtained by resampling the five random seeds with replacement, likewise separate the margin-based from the parameter-count measures (Table 7).

measure	90% $\widehat{SE}$ All [95% CI]	90% $\widehat{SE}$ Common [95% CI]
$1/m_m$	0.7124 [0.7082, 0.7169]	0.7225 [0.7178, 0.7295]
$\sqrt{n_q}/m_m$	0.7144 [0.7099, 0.7183]	0.7274 [0.7216, 0.7342]
n_p/m_m	0.7256 [0.7209, 0.7300]	0.7531 [0.7468, 0.7594]
n_p	0.7724 [0.7637, 0.7809]	0.7724 [0.7637, 0.7809]
$n_p\sqrt{n_q}$	0.7724 [0.7638, 0.7810]	0.7724 [0.7634, 0.7810]
composite ($c=0.075$)	0.7117 [0.7075, 0.7159]	0.7212 [0.7156, 0.7275]
$n_p \ln(n_p/m_m)$	0.7321 [0.7275, 0.7367]	0.7699 [0.7609, 0.7778]

Table 6: Bootstrap 95% confidence intervals for the 90th-percentile sign-error, resampling environments.

measure	$\mathcal{K}(\mu)$ [95% CI over seeds]
$1/m_m$	4.565×10^{-3} [3.73, 5.18] $\times 10^{-3}$
$\sqrt{n_q}/m_m$	3.543×10^{-3} [2.87, 4.33] $\times 10^{-3}$
n_p/m_m	4.683×10^{-4} [2.00, 7.96] $\times 10^{-4}$
$n_p\sqrt{n_q}$	1.564×10^{-4} [1.56, 5.36] $\times 10^{-4}$
n_p	1.932×10^{-4} [1.93, 5.25] $\times 10^{-4}$
composite ($c=0.075$)	3.316×10^{-3} [2.69, 4.24] $\times 10^{-3}$
$\hat{R}_{\gamma=0.1}$	1.288×10^{-2} [1.17, 1.54] $\times 10^{-2}$

Table 7: Conditional independence score $\mathcal{K}$ with 95% bootstrap confidence intervals, resampling the five random seeds with replacement (40 resamples). With only five seeds the intervals are coarse; the bootstrap medians preserve the ranking of the point estimates, the intervals cleanly separate the margin-based measures from the parameter-count measures, and place $\hat{R}_{\gamma=0.1}$ above $1/m_m$. For n_p and $n_p\sqrt{n_q}$ no resample falls below the full-grid value of the bootstrap run, which therefore sits at the lower limit.

H Sign-Error Empirical Cumulative Distributions

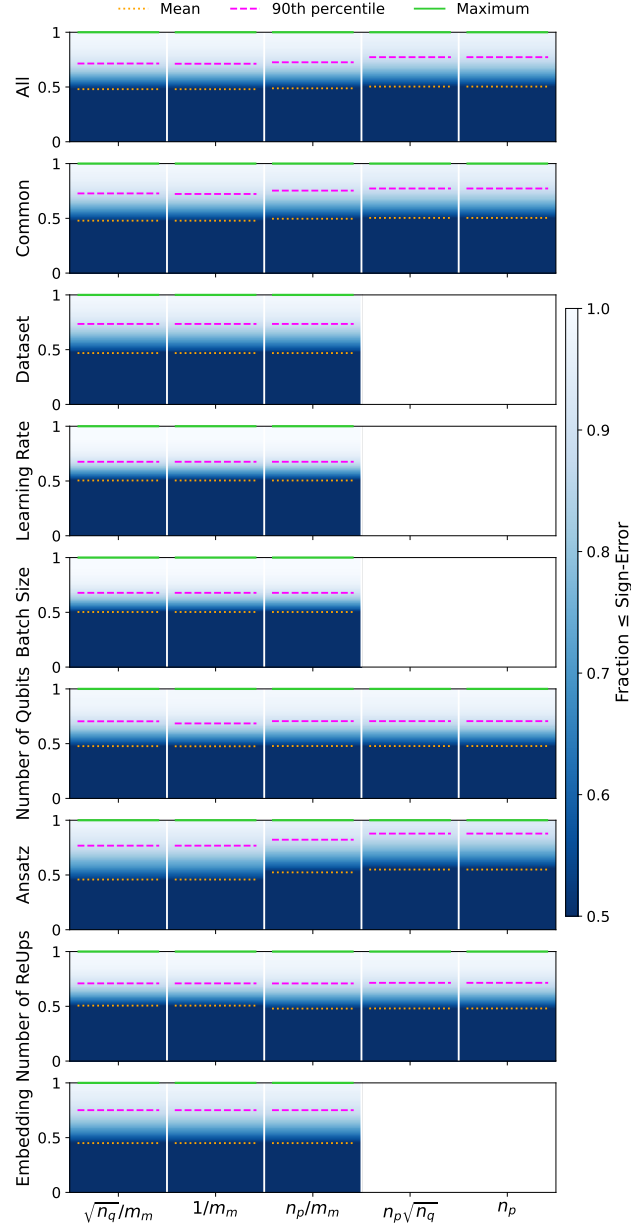


Figure 6: Empirical cumulative distributions of sign-error for the different complexity measures considered in this study. The mean, 90th-percentile, and maximum values are indicated with horizontal lines.